مهندسی صنایع و مدیریت شریف، (تابستان ۱۳۹۶)
دوره‌ی ۱-۳۳، شماره‌ی ۱/۱، ص. ۲۱-۲۸

Original Article

ارائه‌ی الگوریتمی به‌منظور خوشه‌بندی صفحات وب براساس محتوا و لینک

محمد فتحیان * **(استاد)**
امیرمحسن کریمی‌مجد (دانشجوی دکتری)
دانشکده‌ی مهندسی صنایع، دانشگاه علم و صنعت ایران

وجود یک موتور جست‌وجوی کارا می‌تواند سبب افزایش رضایت کاربران از خدمات تحت وب باشد. چالش اصلی موتورهای جست‌وجو، انتخاب مناسب‌ترین صفحات در مواجهه با پرسش‌های چندوجهی کاربران است. «خوشه‌بندی صفحات براساس محتوا و لینک» رویکردی است که برای حل چنین مسائلی در ادبیات پیشنهاد شده است. در این نوشتار، بر یکی از الگوریتم‌های موجود، به‌نام CohsMix [1]، تمرکز شده و این الگوریتم برای ارتقای کیفیت پاسخ‌ها و افزایش سرعت حل بهبود داده شده است. تعیین نقطه‌ی شروع مناسب، استفاده از خواص شبکه‌های پیچیده به‌منظور ساده‌سازی محاسبات، و محاسبه‌ی مقدار واقعی انحراف استاندارد از جمله تغییرات پیشنهادی برای بهبود الگوریتم است. نتایج تجربی نشان می‌دهد که الگوریتم بهبودیافته، کیفیت جواب‌ها را ارتقا داده و باعث افزایش سرعت حل می‌شود. همچنین، به‌عنوان مطالعه‌ی موردی، داده‌های مربوط به وبلاگ‌های فارسی استخراج و الگوریتم بهبودیافته روی این داده‌ها اجرا خواهد شد.

واژگان کلیدی: خوشه‌بندی، تجارت الکترونیکی، محتوا، لینک، موتور جست‌وجو، شبکه‌های پیچیده.

fathian@iust.ac.ir
karimimajd@iust.ac.ir

۱. مقدمه

در دهه‌ی اخیر شاهد رشد و توسعه‌ی چشمگیری در حوزه‌ی تجارت الکترونیکی بوده‌ایم. در این حوزه، اگر چه فروریختن مرزهای جغرافیایی و دسترسی آسان و حتی رایگان به منابع مختلف موجب اقبال کاربران به خدمات الکترونیکی شده، اما سرعت و دقت در دسترسی به منابع مورد نظر در میان انبوه صفحات وب نزد آن‌ها اهمیت ویژه‌یی دارد. به‌عبارت دیگر، یکی از عوامل مؤثر برای تأمین رضایت کاربران از خدمات تحت وب، وجود ابزار مناسبی برای یافتن منابع مورد نظر است. بدین‌ترتیب، برای یک پایگاه تحت وب، برخورداری از یک موتور جست‌وجوی کارا می‌تواند حرکتی در راستای جذب مشتریان و افزایش رضایت‌مندی آنها باشد.

مسئله‌یی که موتورهای جست‌وجو عمدتاً با آن روبرو هستند این است که خروجی آن‌ها در مواجهه با پرسش‌های[2] چندوجهی کاربران، طیف گسترده و پراکنده‌یی از صفحات خواهد بود و انتخاب مناسب‌ترین و مرتبط‌ترین صفحات کار دشواری است. به عبارت ساده‌تر احتمال این که خروجی موتور جست‌وجو در نظر کاربران نامرتبط جلوه کند، بسیار زیاد است. به‌عنوان مثال واژه‌ی «شیر» حداقل در سه معنا کاربرد دارد: ۱. حیوانی وحشی؛ ۲. شیرگاو؛ ۳. شیرآب. حال زمانی که کاربری شیر به‌معنای شیر آب را مورد جست‌وجو قرار دهد و نتایج جست‌وجو حول دو کاربرد دیگر (که در مقایسه با شیر آب بیشتر مورد استفاده قرار می‌گیرند) ارائه شود، موتور جست‌وجو نتوانسته رضایت کاربر را تأمین کند. تقریباً از ابتدای ظهور وب، برای حل چنین معضلی رویکرد «خوشه‌بندی[3] موضوعی صفحات» پیشنهاد شد. بدین‌ترتیب برای هر مفهوم متفاوت، یک خوشه و از هر خوشه نیز یک نماینده در نتایج جست‌وجو قابل رؤیت خواهد بود. این رویکرد در ابتدا مورد استقبال واقع شد، اما در ادامه مشخص شد که کارایی لازم را ندارد؛ زیرا قادر نبود گروه‌های متنوعی را که در ارتباط با کاربردهای مختلف واژگان چندبعدی شکل گرفته‌اند از یکدیگر تمیز دهد.

در سال ۲۰۱۰ الگوریتمی به‌نام CohsMix معرفی شد[1] که خوشه‌بندی صفحات را براساس محتوا[4] و لینک[5] میان آن‌ها انجام می‌داد. برتری این روش نسبت به پژوهش‌های انجام شده تا قبل از سال ۲۰۱۰ [2و3] این است که آن‌ها با در نظرگرفتن مدل شکل‌گیری لینک‌های میان صفحات، امکان استخراج الگوهای واقعی میان صفحات را مهیا کردند. همچنین در نظرگرفتن توزیع آماری مناسب برای مقادیر مربوط به ویژگی‌های صفحات، به مدل آن‌ها ــ در مقایسه با مدل‌های مبتنی بر فاصله[2و3] ــ انعطاف‌پذیری بیشتری می‌دهد. علی‌رغم چنین قابلیت‌های مهمی، حجم بالای محاسبات لازم برای اجرای این الگوریتم به زمان‌گیر بودن آن منجر شده است.

در این مقاله ابتدا بر مشکل نحوه‌ی انجام محاسبات الگوریتم CohsMix و ارائه‌ی الگوریتمی برای بهبود آن تمرکز خواهد شد. ایده‌ی اصلی الگوریتم بهبودیافته در مقایسه با الگوریتم پایه، بهره‌گیری بیشتر از ویژگی‌های ماتریس مجاورت و اطلاعات

* نویسنده مسئول
تاریخ: دریافت ۱۳۹۳/۳/۲۷، اصلاحیه ۱۳۹۴/۴/۷، پذیرش ۱۳۹۴/۵/۷.

مربوط به توزیع احتمالات خصیصه‌ها در راستای کاهش زمان حل مسئله و افزایش کیفیت پاسخ‌هاست. به‌عنوان مطالعه‌ی موردی، برای اولین بار داده‌های وبلاگ‌های فارسی را استخراج کرده و به خوشه‌بندی آن‌ها به‌منظور فراهم آوردن بستری برای ارتقای خروجی موتورهای جست‌وجوگر خواهیم پرداخت.

در بخش بعدی به پیشینه‌ی پژوهش می‌پردازیم و در بخش ۳ مدل پیشنهادی تشریح خواهد شد. بخش ۴ نتایج عددی حاصل از آزمایش‌های مختلف روی مجموعه داده‌های موجود در ادبیات را ارائه می‌دهد. در بخش ۵ نتیجه‌گیری کلی ارائه خواهد شد.

۲. پیشینه‌ی پژوهش

در این بخش ابتدا پژوهش‌های انجام‌شده در حوزه‌ی خوشه‌بندی صفحات وب به‌منظور ارتقای عملکرد موتورهای جست‌وجو مرور خواهد شد. سپس خلاصه‌وار توصیفی از الگوریتم CohsMix، که اساس مقاله‌ی حاضر است، ارائه می‌شود.

۱.۲. خوشه‌بندی صفحات وب

خروجی یک موتور جست‌وجو، علاوه بر جامعیت، باید توانایی پاسخ‌گویی مناسب به درخواست کاربر را داشته باشد. از این رو، هیرست و پدرسن[۲] در سال ۱۹۹۶ ایده‌ی خوشه‌بندی صفحات وب را براساس موضوعات، برای سازمان‌دهی نتایج جست‌وجو ارائه دادند. سپس زمیر و اتزیونی (۱۹۹۸)[۳] این مسئله را مورد توجه قرار دادند. دستاورد این پژوهش‌ها، با نارسایی‌هایی همراه بود، اما توانست پنجره‌ی نوینی برای ارتقای خدمات وب به روی پژوهش‌گران بگشاید.

صرف پرداختن به موضوعات کلی که در صفحات وب مطرح می‌شود، به این دلیل که رفتارهای اجتماعی نویسندگان صفحات را در نظر نمی‌گیرد، چندان نمی‌تواند گویای گروه‌بندی واقعی و مفید برای سازمان‌دهی نتایج باشد. به‌عبارت دیگر، تهیه‌کنندگان محتوای صفحات مختلف اغلب در حوزه‌ی خود با افراد شبیه به خود رابطه دارند و از لینک آن‌ها در صفحات‌شان استفاده می‌کنند. بنابراین محتوایی که در یک حوزه تهیه شده می‌تواند نشأت‌گرفته از دیدگاه‌های مختلف باشد و هر دیدگاه ممکن است الگویی خاص و کاربرانی خاص داشته باشد. از این رو در سال ۲۰۰۲ هی و همکاران،[۵] بدون در نظرگرفتن محتوای صفحات، خوشه‌بندی آنها را براساس لینک‌های میان صفحات انجام داده‌اند. روش دیگری نیز توسط بکرمن و همکاران (۲۰۰۶)[۶] ارائه شد که از یک روش چندعاملی[۶] برای خوشه‌بندی صفحات استفاده می‌کرد و مبنای شباهت در آن بر پایه‌ی مسیری است که میان هر دو صفحه وجود دارد.

بدین‌ترتیب می‌توان روش‌های بالا را در دو دسته جای داد: ۱. روش‌های مبتنی بر محتوا و موضوع صفحات؛ ۲. روش‌هایی که لینک‌های میان صفحات را عامل تشکیل گروه‌ها می‌دانند. در این میان، به نظر می‌رسد گروه‌بندی واقعی صفحات معطوف به الگوهایی است که در سایه‌ی در نظر گرفتن هم‌زمان این دو موضوع (یعنی محتوا و لینک) قابل شناسایی‌اند. چنین رویکردی توسط زنقی و همکاران (۲۰۱۰)[۱] ارائه شد؛ آنها یک الگوریتم مبتنی بر مدل شکل‌گیری گراف صفحات ارائه دادند که از روش معروف مقادیر انتظاری ـ بهینه‌سازی[۷][۷] بهره می‌برد. بدین ترتیب رویکرد آن‌ها نسبت به رویکردهای دیگر (به‌عنوان مثال مدل‌های مبتنی بر فاصله)[۳و۲] توانایی بیشتری در کشف الگوهای واقعی میان صفحات داشت.

اگر مجموعه‌ی صفحات در قالب یک گراف در نظر گرفته شود، هر صفحه با یک گره قابل نمایش است و لینک میان صفحات، یال‌های این گراف خواهد بود. زنقی و همکاران[۱] واژگان مهم و مکرر صفحات را به‌عنوان خصیصه‌های گره‌ها در نظر گرفتند که در واقع بیان‌گر محتوای صفحات است. از این منظر روش آنها را می‌توان در حوزه‌ی خوشه‌بندی گراف‌های خصیصه‌دار[۸] برشمرد. اولین پژوهش در این حوزه را نویل و همکاران[۸] انجام دادند؛ آن‌ها تعداد خصیصه‌های با مقادیر مشابه را به‌عنوان معیار شباهت دو گره در نظر گرفتند. سپس ماتریس مجاورت وزن‌دار جدیدی با استفاده از حاصل‌ضرب این معیار در ماتریس مجاورت گره‌ها تعریف کردند و خوشه‌بندی براساس ماتریس جدید را به‌عنوان روشی که به‌طور همزمان ویژگی‌های ساختاری و خصیصه‌های گره‌ها را لحاظ می‌کند پیشنهاد دادند. روش دیگری نیز در سال ۲۰۰۸ توسط اشتینهاسر و چاولا[۹] ارائه شد که همین رویکرد را دنبال می‌کرد. ژو و همکاران[۱۰] در سال ۲۰۰۹ یک روش مبتنی بر فاصله توسعه دادند. آن‌ها معیار جدیدی برای سنجش فاصله‌ی میان گره‌ها تعریف کردند. این معیار فاصله‌ی گره‌ها را، از نظر ساختار و خصیصه‌ها، بر مبنای فاصله‌ی قدم‌زدن تصادفی[۹] محاسبه می‌کند. در سال ۲۰۱۲ چنگ و همکاران[۱۱] این روش را با استفاده از وزن‌دهی به گره‌ها و ارائه‌ی روشی جدید برای به‌روزکردن مرکز خوشه‌ها بهبود داده‌اند.

رویکرد مهم دیگر، رویکرد مبتنی بر مدل است. یک دسته از روش‌هایی که برمبنای این رویکرد شکل گرفته‌اند، از مدل احتمالاتی بیزی[۱۰] استفاده کرده‌اند.[۱۲و۱۳] به‌عبارت دیگر مسئله‌ی خوشه‌بندی گراف را به یک مسئله‌ی استنباط احتمالاتی تبدیل کرده‌اند.

۲.۲. مروری بر الگوریتم CohsMix

اساس مدل پیشنهادی در نوشتار حاضر الگوریتم موسوم به CohsMix است (شکل ۱)، که در سال ۲۰۱۰ توسط زنقی و همکاران[۱] ارائه شد و روشی مبتنی بر استنباط آماری است. الگوریتم CohsMix گونه‌یی از روش‌های تکرار

الگوریتم CohsMix
ورودی: ماتریس مجاورت X و ماتریس خصیصه ها Y
۱ تعیین مقادیر اولیه برای پارامترها (α ها، π ها، τ_{iq} ها و σ ها) به طور تصادفی
۲ در حالی که مقادیر همگرا نشده اند انجام بده
۳ برای هر گره $i \in \{1, 2, ..., N\}$ انجام بده
۴ برای هر خوشه $q \in \{2,3,...,Q\}$ انجام بده
۵ مقادیر τ_{iq} را با استفاده از رابطه (۱) به روز کن
۶ مقدار τ_{iq} مربوط به هر گره را بر مجموع τ_{iq} های آن گره تقسیم
۷ برای هر خوشه $q \in \{2, 3, ..., Q\}$ انجام بده
۸ مقدار α ها را با استفاده از رابطه (۲) به روز کن
۹ برای هر خوشه $q \in \{2, 3, ..., Q\}$ انجام بده
۱۰ مقادیر π ها را با استفاده از رابطه (۳) به روز کن
۱۱ مقادیر μ ها را با استفاده از رابطه (۴) به روز کن
۱۲ یک واحد به m اضافه کن
خروجی: مقادیر نهایی برای τ_{iq} ها و سایر پارامترها (Θ)

شکل ۱. الگوریتم CohsMix.

شونده‌ی[۷] EM است که در آن زنقی و همکاران مفروضات مدل ساده‌ی گراف کاملاً تصادفی اردوس و رنیی[۱۴] را به‌عنوان مفروضات مدل خود پذیرفتند. به‌عبارت دیگر آن‌ها فرض کردند که یال‌های یک شبکه به‌طور تصادفی و با احتمال مشخص ایجاد شده است. مشخصه‌ی دیگر این الگوریتم این است که توزیع احتمالی خصیصه‌های گره‌های هم‌خوشه را نرمال در نظر می‌گیرد.

درمورد گراف یا شبکه G فرض کنید که X_{ij} نمایانگر درایه‌های ماتریس مجاورت گره‌ها و Y_{ik} درایه‌های ماتریس مربوط به مقادیر K ویژگی $(k = ۱, ۲, ..., K)$ هریک از گره‌ها باشد. الگوریتم CohsMix (شکل ۱) با در نظرگرفتن ماتریس‌های X و Y تلاش دارد همه‌ی N گره موجود را به Q خوشه تخصیص دهد. به‌عبارت دیگر این روش از توزیع احتمالاتی X و Y برای تخمین میزان وابستگی هر گره به هریک از Q خوشه استفاده می‌کند. این میزان وابستگی با τ_{iq} نشان داده می‌شود که در آن $q = ۲, ۳, ..., Q$. در ابتدا مقادیر اولیه‌ی τ_{iq}ها به‌طور تصادفی تعیین می‌شود اما در ادامه، مقادیر به‌دست آمده در هر مرحله از الگوریتم روش CohsMix توسط رابطه‌ی ۱ به‌روز می‌شود:

$$\tau_{iq}^{(m+۱)} = \alpha_q^{(m)} \prod_{j\neq i}\prod_{l} [(\pi_{ql}^{(m)})^{X_{ij}} (۱-\pi_{ql}^{(m)})^{۱-X_{ij}}]^{\tau_{iq}^{(m)}} \times \prod_{k=۱}^{K} [\exp(\frac{۱}{۲\sigma_q^{(m)^۲}})(Y_{ik}-\mu_{qk}^{(m)})^۲] \quad (۱)$$

که در آن:

$$\alpha_q^{(m+۱)} = \frac{۱}{N}\sum_{i=۱}^{N} \tau_{iq}^{(m+۱)}, \quad (۲)$$

$$\pi_{ql}^{(m+۱)} = \frac{\sum_{i\neq j} \tau_{iq}^{(m+۱)} \tau_{jl}^{(m+۱)} X_{ij}}{\sum_{i\neq j} \tau_{iq}^{(m+۱)} \tau_{jl}^{(m+۱)}}, \quad (۳)$$

$$\mu_q^{(m+۱)} = \frac{\sum_i \tau_{iq}^{(m+۱)} Y_i}{\sum_i \tau_{iq}^{(m+۱)}}, \quad (۴)$$

پیچیدگی زمانی این الگوریتم از مرتبه‌ی $O(N^۲)$ است. تعداد بهینه‌ی خوشه‌ها (یا همان Q) با اجرای الگوریتم برای مقادیر متفاوت Q و باتوجه به بیشینه‌ی مقدار به دست آمده برای هر Q با استفاده از معیار ICL[۱۱][۱۵] به دست می‌آید:

$$\begin{aligned}\mathrm{ICL}(Q) = \max_{\Theta} \log L(X, Y, Z; \Theta, Q) \\ - \frac{۱}{۲}Q(Q-۱)\log(\frac{N(N-۱)}{۲}) - \frac{Q-۱}{۲}\log(N) \\ - K(K-۱)\log(\frac{N(N-۱)}{۲}) + KQ\log(\frac{N(N-۱)}{۲})\end{aligned} \quad (۵)$$

جمله‌ی اول معیار ICL بیشینه‌ی لگاریتم راستی‌نمایی ماتریس‌های ورودی، و Z متغیر صفر و ۱ است که نمایانگر تخصیص یا عدم‌تخصیص یک گره به هریک از Q خوشه است.

۳. الگوریتم پیشنهادی

الگوریتم CohsMix با تعداد اندکی تکرار به حالت بهینه‌ی محلی خود می‌رسد و سازوکاری برای فرار از بهینه‌ی محلی ندارد. بنابراین برای رسیدن به یک جواب قابل قبول لازم است چندین بار الگوریتم اجرا شود و بهترین اجرا ثبت شود. اگرچه فقط حدود ۱۰ الی ۱۵ تکرار برای رسیدن به حالت بهینه‌ی محلی لازم است، محاسبات مختص هر تکرار ــ به‌ویژه استفاده از رابطه‌ی به‌روزکننده‌ی τ_{iq}ها یا همان رابطه‌ی ۱ ــ برای مسائل با اندازه‌ی متوسط و بزرگ زمان‌گیر است. از این رو، یافتن نقطه‌ی شروع مناسب و استفاده از روش‌های جایگزین برای کاهش محاسبات می‌تواند رویکردی مؤثر در افزایش کارایی الگوریتم CohsMix باشد. راهکار پیشنهادی در اینجا این است که خروجی الگوریتم خوشه‌بندی K-means به‌عنوان نقطه‌ی شروع نسبتاً مناسبی برای الگوریتم CohsMix تعریف شود. بدین‌ترتیب τ_{iq}ها و α_qهای اولیه به دست می‌آید.

پیش از پرداختن به ادامه‌ی مراحل، ذکر این نکته ضروری است که زنقی و همکاران[۱] در الگوریتم خود مقدار انحراف استاندارد توزیع نرمال مربوط به خصیصه‌های هر خوشه را برابر مقدار ثابت ۱ در نظرگرفته‌اند. چنین فرضی می‌تواند واقعی‌بودن نتایج را خدشه‌دار کند. از این رو، با توجه به پارامترها و متغیرهای تعریف شده در الگوریتم CohsMix، رابطه‌یی برای محاسبه‌ی انحراف استاندارد توزیع نرمال درباره‌ی خصیصه‌های هر خوشه به دست آورده‌ایم. رابطه‌ی ۶ رابطه‌ی مورد نظر است که در هر تکرار به روز می‌شود.

پس از یافتن مقادیر اولیه‌ی τ_{iq}ها و α_qها براساس خروجی الگوریتم K-means، محاسبه‌ی π_{ql}ها از رابطه‌ی ۳، و محاسبه‌ی پارامترهای توزیع نرمال هر خوشه طبق روابط ۴ و ۶ ممکن خواهد بود. نتیجه‌ی استفاده از رابطه‌ی ۶، به دست آوردن یک انحراف استاندارد جداگانه برای هر خوشه خواهد بود.

$$\sigma_q^{(m+۱)} = \sqrt{\frac{\sum_i \tau_{iq}^{(m+۱)} (Y_i - \mu_q^{(m+۱)})^T (Y_i - \mu_q^{(m+۱)})}{\sum_i \tau_{iq}^{(m+۱)}}}, \quad (۶)$$

در پایان الگوریتم CohsMix با توجه به مقادیر جدید اجرا خواهد شد. همچنین از آنجا که عناصر ماتریس X_{ij} مقادیر صفر یا ۱ را اتخاذ می‌کند، می‌توان جمله دوم:

$$A = \prod_{j\neq i}\prod_{l} [(\pi_{ql}^{(m)})^{X_{ij}} (۱-\pi_{ql}^{(m)})^{۱-X_{ij}}]^{\tau_{iq}^{(m)}}, \quad (۷)$$

را که حاصل‌ضرب عبارت‌های:

$$P = [(\pi_{ql}^{(m)})^{X_{ij}} (۱-\pi_{ql}^{(m)})^{۱-X_{ij}}]^{\tau_{iq}^{(m)}}, \quad (۸)$$

است به دو بخش $p_۱$ و $p_۲$ تبدیل کرد که سریع‌تر قابل محاسبه‌اند. در واقع، اگر گره‌های i و j با یکدیگر رابطه داشته باشند (یعنی $X_{ij} = ۱$) آنگاه:

$$P_۱ = \pi_{ql}^{(m)^{\tau_{iq}^{(m)}}}, \quad (۹)$$

و در غیر این صورت:

$$P_۲ = (۱-\pi_{ql}^{(m)})^{\tau_{iq}^{(m)}}, \quad (۱۰)$$

استفاده از دو رابطه‌ی ۹ و ۱۰ مستلزم این است که مقادیر π_{ql} صفر یا ۱ نباشد که این فرض نیز دور از واقعیت نیست. بدین‌ترتیب نوآوری‌های مدل پیشنهادی عبارت‌اند از: ۱. جلوگیری از اجراهای مکرر با ایجاد جواب‌های اولیه‌ی مناسب حاصل به‌کارگیری الگوریتم K-means؛ ۲. کاهش زمان اجرای الگوریتم مورد نظر با تعریف روابط ۹ و ۱۰؛ ۳. افزایش کیفیت جواب‌ها با استفاده از محاسبه‌ی انحراف استاندارد واقعی برای هر خوشه (شکل ۲).

الگوریتم جدید
ورودی: ماتریس مجاورت X و ماتریس خصیصه ها Y
۱ خوشه بندی ماتریس خصیصه ها با استفاده از الگوریتم K-means
۲ محاسبه مقادیر (α ها، π ها، τ_{iq} ها و σ ها) با توجه به خروجی گام قبل
۳ در حالی که مقادیر همگرا نشده اند انجام بده
۴ برای هر گره $i \in \{1, 2, ..., N\}$ انجام بده
۵ برای هر خوشه $q \in \{2, 3, ..., Q\}$ انجام بده
۶ مقادیر τ_{iq} را با استفاده از روابط (۱)، (۹) و (۱۰) به روز کن
۷ مقدار τ_{iq} مربوط به هر گره را بر مجموع τ_{iq} های آن گره تقسیم کن
۸ برای هر خوشه $q \in \{2, 3, ..., Q\}$ انجام بده
۹ مقدار α ها را با استفاده از رابطه (۲) به روز کن
۱۰ برای هر خوشه $l \in \{2, 3, ..., Q\}$ انجام بده
۱۱ مقادیر π ها را با استفاده از رابطه (۳) به روز کن
۱۲ مقادیر μ ها را با استفاده از رابطه (۴) به روز کن
۱۳ مقادیر μ ها را با استفاده از رابطه (۶) به روز کن
۱۴ یک واحد به m اضافه کن
خروجی: مقادیر نهایی برای τ_{iq} ها و سایر پارامترها (Θ)

شکل ۲. الگوریتم جدید.

۴. نتایج عددی

در این بخش ابتدا داده‌های مورد استفاده معرفی می‌شوند. سپس نتایج اجرای دو روش (روش اصلی CohsMix و روش پیشنهادی) روی مجموعه‌داده‌های مختلف ارائه و تأثیر استفاده از نوآوری‌های روش پیشنهادی برای کاهش زمان اجرای الگوریتم و افزایش کیفیت جواب‌ها بررسی می‌شود. لازم به ذکر است از آنجا که الگوریتم CohsMix با در نظرگرفتن نقاط شروع تصادفی اجرا می‌شود، برای این که مقایسه‌ی منصفانه‌یی میان نتایج دو الگوریتم داشته باشیم، لازم است الگوریتم CohsMix (الگوریتم پایه) بیشتر از یک بار اجرا شود و میانگین نتایج به دست آمده از ۱۰ بار اجرای این الگوریتم در مقایسه و ارزیابی در نظرگرفته شود.

۱.۴. آزمایش روی مجموعه‌داده‌های مصنوعی

یکی از روش‌های مقایسه‌ی الگوریتم‌ها استفاده از مجموعه‌داده‌یی است که اطلاعات کافی نسبت به خوشه‌های موجود در آن در دسترس باشد. با داشتن چنین داده‌هایی می‌توان تعیین کرد که خروجی کدام‌یک از الگوریتم‌ها بیشتر شبیه آن چیزی است که در مجموعه‌داده وجود دارد. بدین‌منظور در بخش بعد چنین مجموعه‌داده‌یی را معرفی می‌کنیم و سپس به بررسی نتایج خواهیم پرداخت.

۱.۱.۴. مجموعه داده‌های مصنوعی

مجموعه‌داده‌های مورد استفاده در این نوشتار از مطالعه‌ی زنقی و همکاران[۱] انتخاب شده است. این مجموعه‌داده‌ها، ماتریس‌های شبیه‌سازی شده توسط رایانه است که با در نظرگرفتن پارامترهای زیر ایجاد شده‌اند.

شبیه‌سازی ماتریس مجاورت X برای این مجموعه‌داده‌ها با به‌کارگیری فرایند ایجاد گراف تصادفی و نیز توزیع آماری خصیصه‌ها، با استناد به قضیه‌ی حد مرکزی

جدول ۱. اطلاعات مربوط به پارامترهای مجموعه‌داده‌ها.

مسائل	پارامترها				مسائل	پارامترها			
	N	Q	$d(\mu_q, \mu_l)$	σ		N	Q	$d(\mu_q, \mu_l)$	σ
۱	۱۰۰	۳	۳	۲	۱۱	۱۵۰	۴	۱	۲
۲	۱۰۰	۴	۳	۲	۱۲	۱۵۰	۴	۲	۲
۳	۱۰۰	۵	۳	۲	۱۳	۱۵۰	۴	۳	۲
۴	۱۰۰	۶	۳	۲	۱۴	۱۵۰	۴	۴	۲
۵	۱۰۰	۷	۳	۲	۱۵	۱۵۰	۴	۵	۲
۶	۱۵۰	۳	۳	۲	۱۶	۱۵۰	۴	۵	۱/۵
۷	۲۰۰	۳	۳	۲	۱۷	۱۵۰	۴	۵	۲/۵
۸	۲۵۰	۳	۳	۲	۱۸	۱۵۰	۴	۵	۳
۹	۳۵۰	۳	۳	۲	۱۹	۱۵۰	۴	۵	۳/۵
۱۰	۵۰۰	۳	۳	۲					

جدول ۲. نتایج مربوط به زمان اجرای دو الگوریتم روی مسائل با تعداد گره متفاوت برحسب ثانیه.

روش‌ها	مسائل					
	۱	۶	۷	۸	۹	۱۰
CohsMix	۲/۱۲	۴/۷۴	۸/۴	۱۲/۹	۲۵/۰۸	۵۰/۵۹
الگوریتم جدید	۱/۵۳	۳/۳۲	۵/۸۵	۹/۱۹	۱۷/۸۱	۳۴/۸۹
اختلاف نسبی	۰/۲۸	۰/۳	۰/۳	۰/۲۹	۰/۲۹	۰/۳۱

برای اعداد بزرگ، توزیع نرمال در نظرگرفته شده است. در این جدول، N تعداد گره‌ها، Q بیانگر تعداد خوشه‌های ازپیش تعیین‌شده برای تولید ماتریس‌ها و $d(\mu_l, \mu_q)$ تفاضل میان میانگین خصیصه‌های دو خوشه است. معیار فاصله برای در نظر گرفتن این تفاضل فاصله‌ی اقلیدسی لحاظ شده است. ستون σ اشاره به مقدار انحراف استاندارد میان مقادیر خصیصه‌ها دارد. همچنین پارامتر دیگری در ایجاد مجموعه‌داده‌ها کاربرد داشته که تفاضل میان: «احتمال برقراری ارتباط میان یک گره و گره‌های هم‌خوشه» با «احتمال برقراری ارتباط میان آن گره و گره‌های ناهم‌خوشه» را نشان می‌دهد. از آنجا که این پارامتر برای همه‌ی مجموعه‌داده‌ها یکسان و برابر ۰/۲ است در جدول مورد اشاره قرار نگرفته است. خصیصه‌ها مقادیری گویا دارند و تعداد آن‌ها برای هر گره در همه‌ی مجموعه‌داده‌ها برابر ۳ است. در جدول ۱ اطلاعات مربوط به مجموعه‌داده‌های مختلف ارائه شده است.

۲.۱.۴. مقایسه‌ی روش‌ها از نظر زمان اجرا

در این بخش ابتدا دو الگوریتم (الگوریتم CohsMix و الگوریتم جدید پیشنهادی) از نظر زمان اجرا برحسب ثانیه روی مسائلی که تعداد خوشه‌ی متفاوت دارند مقایسه می‌شود. انتظار می‌رود نوآوری‌های مربوط به کاهش حجم محاسبات بتواند بهبود قابل‌توجهی به هنگام رشد مسئله نشان دهد. در ادامه، زمان اجرای دو الگوریتم روی مسائل متفاوت از نظر تعداد خوشه $d(\mu_q, \mu_l)$ و انحراف استاندارد با هم مقایسه خواهد شد. لازم به یادآوری است که نتایج ارائه شده برای الگوریتم CohsMix میانگین نتایج حاصل از ۱۰ بار اجرای این الگوریتم است.

نتایج مربوط به استفاده از دو روش در جدول ۲ آمده است. طبق این جدول تفاوت قابل ملاحظه‌یی میان زمان اجرای دو روش رؤیت می‌شود. مشخص است که با افزایش تعداد گره‌ها (از ۱۰۰ به ۵۰۰ گره طبق جدول ۱)، این اختلاف نیز بیشتر می‌شود. از سوی دیگر می‌توان نتیجه گرفت که الگوریتم جدید همواره حدود

جدول ۳. نتایج مربوط به زمان اجرای دو الگوریتم روی مسائل با تعداد خوشه متفاوت برحسب ثانیه.

روش‌ها	مسائل				
	۱	۲	۳	۴	۵
CohsMix	۲٫۱۲	۳٫۴۴	۵٫۰۵	۶٫۸۸	۹٫۱۱
الگوریتم جدید	۱٫۵۳	۲٫۴	۳٫۴۷	۴٫۶۶	۶٫۲۴
اختلاف نسبی	۰٫۲۸	۰٫۳	۰٫۳۱	۰٫۳۲	۰٫۳۲

جدول ۴. نتایج مربوط به زمان اجرای دو الگوریتم روی مسائل با $d(\mu_q, \mu_l)$ متفاوت برحسب ثانیه.

روش‌ها	مسائل				
	۱۱	۱۲	۱۳	۱۴	۱۵
CohsMix	۷٫۷۴	۷٫۷۲	۷٫۷۹	۷٫۶۵	۷٫۶۳
الگوریتم جدید	۵٫۲۹	۵٫۳۳	۵٫۳۱	۵٫۳	۵٫۳
اختلاف نسبی	۰٫۳۲	۰٫۳۱	۰٫۳۲	۰٫۳۱	۰٫۳۱

جدول ۵. نتایج اجرای دو الگوریتم روی مسائل با انحراف استاندارد متفاوت برحسب ثانیه.

روش‌ها	مسائل				
	۱۶	۱۵	۱۷	۱۸	۱۹
CohsMix	۷٫۴۵	۷٫۶۳	۷٫۵۲	۷٫۵۷	۷٫۵
الگوریتم جدید	۵٫۳	۵٫۳	۵٫۳۲	۵٫۳۸	۵٫۲۶
اختلاف نسبی	۰٫۲۹	۰٫۳۱	۰٫۲۹	۰٫۲۹	۰٫۳

۳۰ درصد زمان اجرا را بهبود داده است. چنان‌که اشاره شد، این بهبود در وهله‌ی اول معطوف به کاهش تعداد تکرارهاست که در نتیجه‌ی یافتن نقطه‌ی شروع مناسب با استفاده از الگوریتم خوشه‌بندی صورت می‌پذیرد. در وهله‌ی دوم کاهش زمان انجام هر تکرار الگوریتم که در نتیجه‌ی ساده‌سازی محاسبات مرتبط با به‌روزکردن متغیرها حاصل می‌شود عاملی برای کاهش زمان حل خواهد بود. بدین‌ترتیب تأثیر این تغییرات در مجموعه‌داده‌های مختلف تقریباً مشابه بوده و از نظر زمانی قابل توجه است.

در جدول ۳ زمان اجرای دو الگوریتم، هنگامی‌که تعداد خوشه‌های تعریف‌شده در مجموعه‌داده‌های مصنوعی متفاوت است، ارائه شده است. این جدول نیز بیانگر کاهش زمان اجرا به‌هنگام استفاده از الگوریتم جدید است. همچنین اختلاف نسبی زمان اجرای دو الگوریتم با افزایش تعداد خوشه‌ها (از ۳ به ۷ خوشه طبق جدول ۱) رشد کمی را نشان می‌دهد.

زمان اجرای دو الگوریتم روی مجموعه‌داده‌های متفاوت از نظر $d(\mu_q, \mu_l)$ و انحراف استاندارد به‌ترتیب در جدول‌های ۴ و ۵ ارائه شده است. مشخص است که تفاوت این دو پارامتر، از آنجا که در پیچیدگی زمانی الگوریتم تأثیری ندارند، تفاوت محسوسی در زمان اجرای دو الگوریتم نخواهد داشت. از طرف دیگر، نتایج به دست‌آمده از زمان اجرای الگوریتم جدید در هر دو جدول کم‌تر از روش CohsMix گزارش شده است.

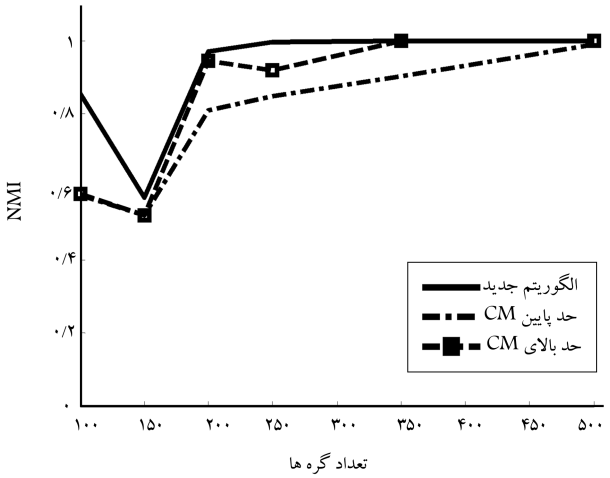
۳.۱.۴. مقایسه‌ی روش‌ها از نظر کیفیت خوشه‌بندی

برای مقایسه‌ی کیفیت پاسخ‌های دو روش نیاز به معیاری است که با توجه به مشخص بودن برچسب هر گره، میزان صحت خوشه‌بندی انجام گرفته توسط هر روش را به شکل کمی بیان کند. برای این مهم، معیار NMI [۱۶] ۱۲ در نظر گرفته می‌شود. محاسبه‌ی این معیار برای دو مجموعه‌ی A و B مطابق رابطه‌ی ۱۱ انجام می‌شود:

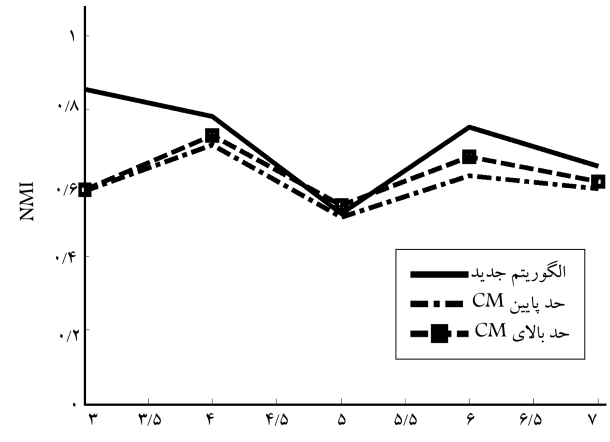
$$I(A,B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B}\frac{C_{ij}\log(C_{ij}N)}{C_{i\bullet}C_{\bullet j}}}{\sum_{i=1}^{C_A}C_{i\bullet}\log(\frac{C_{i\bullet}}{N})\sum_{j=1}^{C_B}C_{\bullet j}\log(\frac{C_{\bullet j}}{N})}, \qquad (۱۱)$$

که در آن C_A و C_B به‌ترتیب نماینده‌ی تعداد خوشه‌ها در دو مجموعه‌ی A و B هستند و $C_{i.}$، $C_{.j}$ و C_{ij} نیز به‌ترتیب بیان‌گر تعداد گره‌ها در خوشه‌ی iام، jام و مشترک بین iام و jام است. هرچه میزان NMI بیشتر باشد خوشه‌بندی با کیفیت بالاتری حاصل شده است. در مسائل مورد بحث در نوشتار حاضر استفاده از معیار NMI امکان‌پذیر خواهد بود، زیرا مجموعه‌داده‌ها توسط رایانه ایجاد شده و اطلاعاتی در این خصوص که کدام‌یک از گره‌ها با هم هم‌خوشه‌اند وجود دارد.

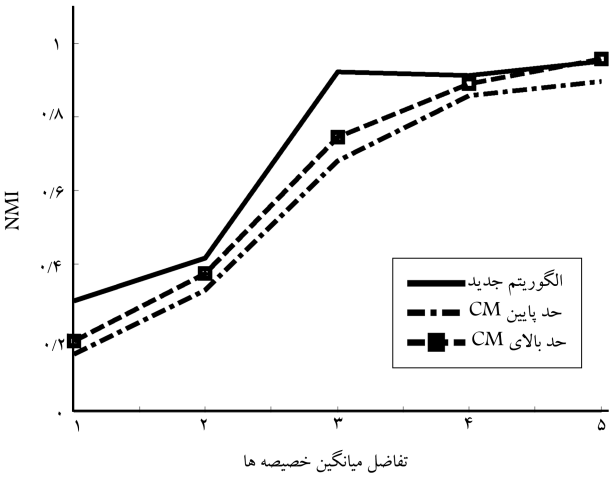
نتایج به دست‌آمده از اجرای دو الگوریتم روی مجموعه‌داده‌های مختلف برحسب معیار NMI در شکل‌های ۳ تا ۶ نشان داده شده است. در این شکل‌ها، الگوریتم جدید با کم‌ترین و بیشترین مقدار به دست‌آمده از ۱۰ بار اجرای الگوریتم CohsMix (که به‌ترتیب با عناوین حد پایین CM و حد بالا CM نشان داده شده‌اند) مقایسه می‌شود.



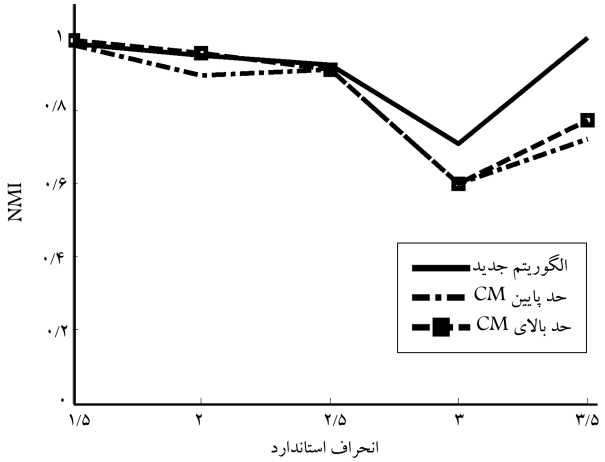
شکل ۳. نتیجه‌ی اجرای دو الگوریتم روی مسائل با تعداد گره متفاوت براساس معیار NMI.



شکل ۴. نتیجه‌ی اجرای دو الگوریتم روی مسائل با تعداد خوشه‌ی متفاوت براساس معیار NMI.



شکل ۵. نتیجه‌ی اجرای الگوریتم‌ها براساس NMI روی مسائل با مقدار متفاوت تفاضل میانگین خصیصه‌ها.



شکل ۶. نتیجه‌ی اجرای دو الگوریتم روی مسائل با مقدار انحراف استاندارد متفاوت براساس معیار NMI.

در شکل ۳ نتیجه‌ی اجرای دو الگوریتم روی مسائل با تعداد گره متفاوت براساس معیار NMI نشان داده شده است. چنان که مشاهده می‌شود در اغلب مسائل، الگوریتم جدید خوشه‌بندی با کیفیت بالاتری را به ارمغان آورده است. از سوی دیگر، با افزایش تعداد گره‌ها و درنتیجه افزایش تعداد اعضای هر خوشه، تقریب‌های بهتری از توزیع نرمال خصیصه‌های اعضای هر خوشه حاصل شده و درنتیجه کیفیت خوشه‌بندی برحسب معیار NMI ارتقاء یافته است.

در شکل ۴ رفتار دو الگوریتم هنگام اجرا روی مجموعه‌داده‌های متفاوت از نظر تعداد خوشه نشان داده شده است. هر دو الگوریتم بر اثر افزایش تعداد خوشه (بدون در نظرگرفتن مقادیر متناظر با تعداد خوشه برابر ۵) سیری نزولی از نظر میزان NMI محاسبه شده دارند. به‌طور کلی انتظار می‌رود برای مسائل با تعداد خوشه‌ی بالاتر از ۴، میزان کیفیت خوشه‌بندی‌های دو الگوریتم از ۰/۸ کم‌تر باشد. از سوی دیگر، چنان که در این شکل مشاهده می‌شود، الگوریتم جدید از نظر کیفیت جواب‌ها در اکثر مسائل برتر است.

در شکل ۵ روندی صعودی برای هر دو الگوریتم هنگام اجرا روی مجموعه‌داده‌های متفاوت از نظر تفاضل میانگین خصیصه‌های خوشه‌ها نشان داده شده است. مشخص است که الگوریتم جدید از نظر کیفیت جواب‌ها در اکثر مسائل برتر از روش پایه است.

البته به نظر می‌رسد زمانی که تفاضل میان میانگین مقادیر مربوط به خصیصه‌ها افزایش می‌یابد، هر دو الگوریتم میل به کشف صحیح خوشه‌ها دارند. علت آن است که افزایش فاصله‌ی میان مقادیر مربوط به خصیصه‌ها به‌معنای افزایش میزان جدایی خوشه‌ها از هم و افزایش توانایی تشخیص از یکدیگر خواهد بود.

در شکل ۶ دو الگوریتم مبتنی بر معیار NMI روی مسائل با مقدار انحراف استاندارد متفاوت مقایسه شده است. لازم به یادآوری است که روش CohsMix همواره مقدار انحراف استاندارد همه‌ی خصیصه‌ها برای همه‌ی خوشه‌ها را برابر ۱ در نظر می‌گیرد. از این رو انتظار می‌رود که با افزایش انحراف استاندارد واقعی خصیصه‌ها، کیفیت خوشه‌بندی در این روش کاهش یابد (شکل ۶). از طرفی انتظار می‌رود الگوریتم جدید که همواره برای همه‌ی خوشه‌ها مقدار انحراف استاندارد را محاسبه می‌کند، کم‌تر متأثر از تغییر انحراف استاندارد باشد. در شکل ۶ فقط زمانی که انحراف استاندارد برابر ۳ شده است میزان NMI محاسبه شده برای الگوریتم جدید کم‌تر از ۰/۹ است. بدین‌ترتیب می‌توان نتیجه گرفت که نوآوری این مقاله در راستای استفاده از رابطه‌ی ۶ برای محاسبه‌ی انحراف استاندارد توزیع نرمال مربوط به خصیصه‌های خوشه‌ها مؤثر بوده است.

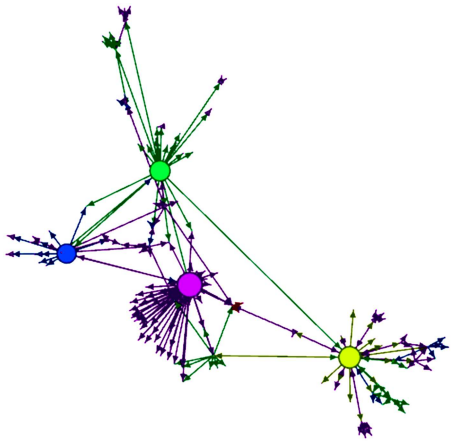
۲.۴. مطالعه‌ی موردی: وبلاگ‌های فارسی

براساس رتبه‌بندی سایت الکسا (Alexa.com) در بهار سال ۲۰۱۴ میلادی، ۳ مورد از ۱۰ سایت پر بازدید توسط کاربران ایرانی مربوط به سایت‌های ارائه‌دهنده‌ی خدمات وبلاگ‌نویسی است. این موضوع به‌خوبی جایگاه وبلاگ‌نویسی در میان کاربران ایرانی و اهمیت مطالعه روی این حوزه را نمایان می‌کند. در این بخش ابتدا به تبیین نحوه‌ی تهیه‌ی داده‌ها، و سپس به بررسی نتایج حاصل از اجرای الگوریتم پیشنهادی روی داده‌های جمع‌آوری شده خواهیم پرداخت.

۱.۲.۴. تهیه‌ی مجموعه‌داده

برای تهیه‌ی داده‌های مذکور لازم است هر دو موضوع ارتباط میان وبلاگ‌ها و محتوای آن‌ها در نظرگرفته شود. به‌عبارت دیگر ما با شبکه‌یی از وبلاگ‌ها مواجهیم و برای جمع‌آوری داده‌ها باید از روش‌های جمع‌آوری داده که در حوزه‌ی شبکه‌های پیچیده مطرح است استفاده شود. از این رو به نظر می‌رسد استفاده از روش خزش در شبکه مناسب باشد. بدین ترتیب ابتدا به‌طور تصادفی یک وبلاگ را از یکی از سایت‌های ارائه‌دهنده‌ی خدمات وبلاگ‌نویسی برمی‌گزینیم. در اینجا سایت پرشین‌بلاگ (Persianblog.ir) مورد توجه واقع شده است. سپس وبلاگ‌هایی که در لیست لینک‌های این وبلاگ (دوستان نویسنده‌ی وبلاگ) هستند انتخاب می‌شود. این نحوه‌ی گزینش درباره‌ی دوستان نویسندگان وبلاگ‌هایی که انتخاب شده‌اند نیز صورت می‌گیرد. لازم به ذکر است که وبلاگ دوستان بعضی از وبلاگ‌های مورد مطالعه در دامنه‌ی سایت‌های دیگر تعریف شده است. در اینجا از در نظر گرفتن چنین وبلاگ‌هایی صرف نظر شده زیرا هدف مطالعه و بررسی وبلاگ‌های تعریف شده در دامنه‌ی پرشین‌بلاگ است.

با لحاظ کردن موارد بالا تعداد ۱۶۲ وبلاگ در نهایت گزینش شد که ۳۹۰ لینک با یکدیگر دارند. محتوای این وبلاگ‌ها، متن پست‌هایی است که در وبلاگ قرار داده شده است. با استفاده از ابزار متن‌کاوی، واژگان به کار رفته و واژگان مهم متون وبلاگ‌ها استخراج شد. تنوع کل واژگان ۱۲۹۰۶ واژه بوده که براساس روش متن‌کاوی ارائه شده[۱۷] تعداد ۸۸ مورد حائز اهمیت تشخیص داده شد. بدین‌ترتیب، خصیصه‌های صفحات در واقع همان واژگان مهم و مقدار آن‌ها همان فراوانی آن‌ها در صفحه‌ی مورد نظر خواهد بود.



شکل ۷. نمایش خوشه‌های مربوط به وبلاگ‌ها.

۲.۲.۴. نتایج مطالعه‌ی موردی

تعداد بهینه‌ی خوشه‌ها براساس خروجی الگوریتم جدید برابر ۴ به دست آمد. در شکل ۷ تعداد ۴ خوشه کشف شده حول مراکز خوشه و نیز صفحاتی که مقادیر به دست آمده برای خصیصه‌های آن‌ها به میانگین مقادیر صفحات هم‌خوشه‌یی آن‌ها شبیه‌ترین است با دوایر بزرگ‌تر نشان داده شده‌اند.

به‌طورکلی این شکل بیانگر تنوع مطالب در میان صفحات جمع‌آوری شده از کاربران پرشین‌بلاگ است. اگر مسئله، یافتن صفحه‌ی مرتبط با پرسش کاربر از میان صفحات موجود (مورد بررسی در داده‌های جمع‌آوری شده) باشد آنگاه ابتدا باید همه‌ی خوشه‌هایی که صفحه‌یی مرتبط با آن پرسش دارند مشخص شود. سپس از هریک از آن‌ها گزینه‌یی انتخاب می‌شود. بدین‌ترتیب در خروجی جست‌وجو، انواع صفحات مرتبط لحاظ شده است.

۵. نتیجه‌گیری

در این نوشتار مسئله‌ی خوشه‌بندی صفحات وب براساس محتوا و لینک بررسی شد. برای حل چنین مسئله‌یی یکی از الگوریتم‌های موجود در ادبیات به کارگرفته شد و تلاش شد تا با اتخاذ نوآوری‌های پیشنهادی، بهبود قابل ملاحظه‌یی ــ چه از لحاظ کیفیت پاسخ‌ها و چه از لحاظ زمان حل ــ ایجاد شود. این نوآوری‌ها در قالب تعیین جواب اولیه‌ی مناسب با به‌کارگیری روش خوشه‌بندی K-means، کاهش حجم محاسبات مربوط به پارامترها با استفاده از ویژگی‌های شبکه‌ی لینک‌ها و به‌کارگیری مقادیر واقعی انحراف استانداردهای هر خوشه بیان شد. نتایج به دست آمده از انجام آزمایش‌های متنوع روی مجموعه‌داده‌های استاندارد حاکی از عملکرد مناسب روش پیشنهادی نسبت به روش CohsMix است. علاوه بر این، مجموعه‌داده‌ی مربوط به وبلاگ‌های فارسی استخراج شد و الگوریتم پیشنهادی بر روی آن داده به‌کارگرفته شد.

پانوشت‌ها

1. covariates on hidden structure using mixture models (CohsMix)
2. query
3. clustering
4. content
5. link
6. multi-agent
7. expectation-maximization
8. attributed graphs
9. random walk
10. bayesian probabilistic model
11. integrated classification likelihood (ICL)
12. normalized mutual index

منابع (References)

1. Zanghi, H., Volant, S. and Ambroise, C. "Clustering based on random graph model embedding vertex features", *Pattern Recogn. Lett*, **31**(9), pp. 830-836 (2010).
2. Weiss, R., Velez, B. and Sheldon, M. "HyPursuit: A hierarchical network search engine that exploits content-link hypertext clustering", *Hypertext'96 Proceedings of the Seventh ACM Conference on Hypertext*, New York, USA, pp. 180-193 (1996).
3. Zamir, O. and Etzioni, O. "Web document clustering: A feasibility demonstration", In: *Proc. 21st Annual Internat. ACM SIGIR Conf. on Research and Development in Information Retrieval*, NewYork, USA, pp. 46-54 (1998).
4. Hearst, M. and Pedersen, J. "Reexamining the cluster hypothesis: Scatter/gather on retrieval results", *19th Annual Internet. ACM SIGIR Conf. Research and Development in Information Retrieval*, New York, USA, pp. 76-84 (1996).
5. He, X., Zha, H., Ding, C.H.Q. and Simon, H.D. "Web document clustering using hyperlink structures", *Computat. Statist. Data Anal.*, **41**(1), pp. 19-45 (2002).
6. Bekkerman, R., Ziberstein, S. and Allan, J "Web page clustering using heuristic search in the web graph", *IJCAI'07 Proceedings of the 20th International Joint Conference on Artifical Intelligence*, New York, USA, pp. 2280-2285 (2006).
7. Dempster, A.P., Laird, N.M. and Rubin, D.B. "Maximum likelihood from incomplete data via the EM algorithm", *J. R. Stat. Soc.*, **39**(1), pp. 1-38 (1977).
8. Neville, J., Adler, M. and Jensen, D. "Clustering relational data using attribute and link information", In *Proceedings of the Text Mining and Link Analysis Workshop, 18th International Joint Conference on Artificial Intelligence*, pp. 689-698 (2003).
9. Steinhaeuser, K. and Chawla, V. "Community detection in a large real-world social network", *Soc. Comput. Beh. Modeling*, pp. 168-175 (2008).

10. Zhou, Y., Cheng, H. and Yu, J.X. "Graph clustering based on structural/attribute similarities", *Proceedings of the VLDB Endowment*, pp. 718-729 (2009).

11. Cheng, H., Zhou, Y., Huang, X. and Yu, J.X. "Clustering large attributed information networks: An efficient incremental computing approach", *Data Min. Knowl. Disc.*, **25**, pp. 450-477 (2012).

12. Kim, M. and Leskovec, J. "Multiplicative attribute graph model of real-world networks", *Internet Math.*, **8**, pp. 113-160 (2012).

13. Xu, Z., Ke, Y., Wang, Y., Cheng, H. and Cheng, J. "A model-based approach to attributed graph clustering", *SIGMOD'12 International Conference on Management of Data*, Scottsdale, Arizona, USA, pp. 505-516 (2012).

14. Erdős, P. and Rényi, A. "On random graphs", *Publicationes 'Mathematicae*, **6**, pp. 290-297 (1959).

15. Biernacki, C., Celeux, G. and Govaert, G. "Assessing a mixture model for clustering with the integrated completed likelihood", *IEEE PAMI*, **22**(7), pp. 719-725 (2000).

16. Danon, L., Diaz-Guilera, A., Duch, J. and Arenas, A. "Comparing community structure identification", *Journal of Statistical Mechanics: Theory and Experiment*, **09**, pp. 1-10 (2005)

17. Banchs, R., *Text Mining with MATLAB*, 1st Edn., chapter 8, Springer (2013).